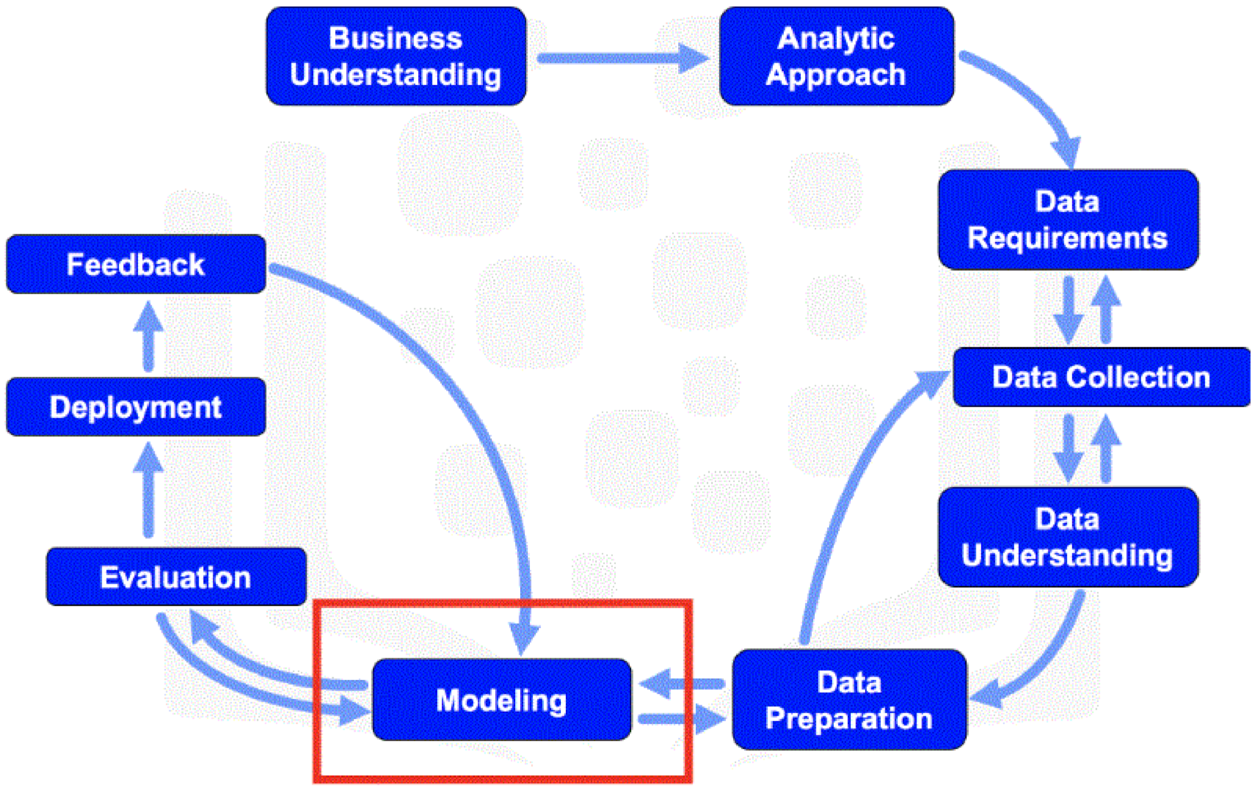
Traffic Accident Severity Prediction

Charles W. Lenfest

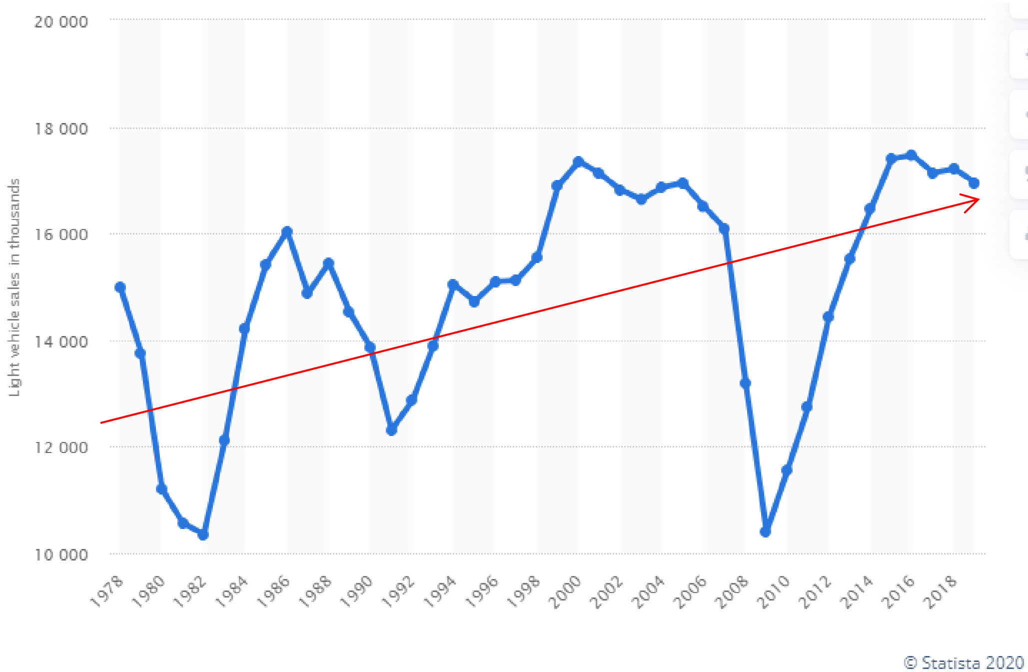
Data Science Methodology



Background

Vehicle sales, while impacted by recessions, have generally increased in tandem with population growth from the 1970s to 2019. This has resulted in dramatic country wide traffic congestion. In 2018 the Seattle metro area ranked as the 6th most congested with respect to traffic in the US. Drivers on average spent over 130 hours stuck in traffic and with limited ability to increase roadways efforts are underway to provide smart solutions to alleviate congestion.

Light vehicle retail sales in the United States from 1978 to 2019



Business Problem

The objective of this project is to develop a Machine Learning model that will predict traffic accident severity based on a data set that includes 36 variables that are correlated to a given accident severity rating. This model could be could potentially be deployed as a smart phone/car app that will allow drivers to monitor traffic conditions while enroute, and adjust the route to their target destination based on the predicted impacted of accidents recorded and warehoused by the SDOT Traffic Management Division, Traffic Records Group.

Population of Interest

This model and app would find a broad population of interest that would span most driver demographics. Given that the average driver wastes over 130+ hours stuck in traffic, it would be advantageous to be able to predict the severity of accidents that contribute to those delays, and allow drivers the ability to alter their route in real time.

Data Pre-processing & Feature Engineering

The data utilized for this study was provided by SDOT Traffic Management Division, Traffic Records Group

Data consists of 37 fields, 36 of which potentially used as Features/Predictors for inputs into the Machine Learning Models. The remaining Severity Code provides the Label/Target field used to train and test the models.

There were numerous issues were discovered with the raw .csv data file, that had to be addressed prior to being used as input for the Machine Learning models. To systematically address these issues we created a Feature Engineering & Pre-processing Operation Schedule

Feature Engineering - Preprocessing Operations Schedule

In [2]:
```
data_wrangle_df = pd.read_csv('D:\ibm_ds_cert\Data Science Capstone Project\DataWrangle.csv',index_col = 0)
data_wrangle_df
```

Out[2]:

#	Feature	FALSE	TRUE	type	X-Check	Unique	Notes
1	X	189339	5334.0	int64	X	23563	mean
2	Y	189339	5334.0	int64	Y	23839	mean
3	OBJECTID	194673	NaN	int64	OBJECTID	194673	drop
4	INCKEY	194673	NaN	int64	INCKEY	194673	drop
5	COLDETKEY	194673	NaN	int64	COLDETKEY	194673	drop
6	REPORTNO	194673	NaN	int64	REPORTNO	194670	drop
7	STATUS	194673	1926.0	int64	STATUS	2	drop
8	ADDRTYPE	129603	65070.0	int64	ADDRTYPE	3	numerical value
9	INTKEY	191996	2677.0	int64	INTKEY	7614	Value = 1 if > 0, 0 if ""
10	LOCATION	191996	2677.0	int64	Location	24102	drop
11	EXCEPTRSNCODE	109862	84811.0	int64	EXCEPTRSNCODE	2	drop
12	EXCEPTRSNDESC	189035	5638.0	int64	EXCEPTRSNDESC	1	drop
13	SEVERITYCODE.1	194673	NaN	int64	SEVERITYCODE.1	2	Label / Target
14	SEVERITYDESC	194673	NaN	int64	SEVERITYDESC	2	encode
15	COLLISIONTYPE	189769	4904.0	int64	COLLISIONTYPE	10	encode
16	PERSONCOUNT	194673	NaN	int64	PERSONCOUNT	47	encode
17	PEDCOUNT	194673	NaN	int64	PEDCOUNT	7	encode
18	PEDCYLCOUNT	194673	NaN	int64	PEDCYLCOUNT	3	encode
19	VEHCOUNT	194673	NaN	int64	VEHCOUNT	13	encode
20	INCDATE	194673	NaN	int64	INCDATE	5985	split into M and DOM
21	INCDTTM	194673	NaN	int64	INCDTTM	162058	Split into 24 int
22	JUNCTIONTYPE	188344	6329.0	int64	JUNCTIONTYPE	7	encode
23	SDOT_COLCODE	194673	NaN	int64	SDOT_COLCODE	39	numerical value
24	SDOT_COLDESC	194673	NaN	int64	SDOT_COLDESC	39	drop
25	INATTENTIONIND	164868	29805.0	int64	INATTENTIONIND	1	Y= 1 , blank = 0
26	UNDERINFL	189789	4884.0	int64	UNDERINFL	4	Chng N = 0, Y = 1
27	WEATHER	189592	5081.0	int64	WEATHER	11	encode
28	ROADCOND	189661	5012.0	int64	ROADCOND	9	encode
29	LIGHTCOND	189503	5170.0	int64	LIGHTCOND	9	encode
30	PEDROWNOTGRNT	190006	4667.0	int64	PEDROWNOTGRNT	1	Y= 1 , blank = 0
31	SDOTCOLNUM	114936	79737.0	int64	SDOTCOLNUM	114932	drop
32	SPEEDING	185340	9333.0	int64	SPEEDING	1	Y = 1,N = 0
33	ST_COLCODE	194655	18.0	int64	ST_COLCODE	115	encode
34	ST_COLDESC	189769	4904.0	int64	ST_COLDESC	62	drop
35	SEGLANEKEY	194673	NaN	int64	SEGLANEKEY	1955	Value = 1 if >0
36	CROSSWALKKEY	194673	NaN	int64	CROSSWALKKEY	2198	Value = 1 if >0
37	HITPARKEDCAR	194673	NaN	int64	HITPARKEDCAR	2	Y = 1,N = 0

Data Pre-processing & Feature Engineering

Examination of the correlation matrix indicates that the strongest correlations relate to the number of vehicles, bicycles, people and pedestrians involved with the accident. Other features included whether pedestrian right of way not granted, address type, driver inattention, under the influence of drugs and alcohol and the precise hour, day of week, month, year in which the accident occurred.

Rank	Feature	Description	Correl to SEVCODE
1	SEVERITYDESC	A detailed description of the severity of the collision	80%
2	PEDCOUNT	The number of pedestrians involved in the collision. This is entered by the state.	24%
3	PEDCYLCOUNT	The number of bicycles involved in the collision. This is entered by the state.	21%
4	INTKEY	Key that corresponds to the intersection associated with a collision	20%
5	SDOT_COLCODE	A code given to the collision by SDOT.	18%
6	PEDROWNOTGRNT	Whether or not the pedestrian right of way was not granted. (Y/N)	18%
7	addrtype_code	Collision address type: • Alley • Block • Intersection	17%
8	PERSONCOUNT	The total number of people involved in the collision	13%
9	Year	Year of accident	8%
10	UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.	7%
11	HofD	Hour of Day of accident	6%
12	INATTENTIONIND	Whether or not collision was due to inattention. (Y/N)	5%
13	Month	Month of accident	3%
14	Day	Day of Week of Accident	2%
15	Y	y coorordinate	2%
16	SPEEDING	Whether or not speeding was a factor in the collision. (Y/N)	1%
17	X	x coordinate	1%
18	VEHCOUNT	The number of vehicles involved in the collision. This is entered by the state.	-4%
19	HITPARKEDCAR	Whether or not the collision involved hitting a parked car. (Y/N)	-4%
20	lightcond_code	Encoded light conditions during the collision.	-11%
21	roadcond_code	Encoded road conditions during the collision.	-11%
22	weather_code	Encoded weather conditions during the collision.	-13%
23	colltype_code	Encoded collision type during the collision.	-14%
24	junctype_code	Encoded junction type during the collision.	-22%

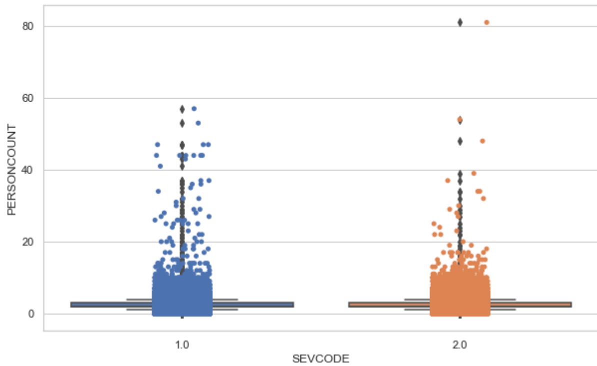
Data Pre-processing & Feature Engineering – Feature Correlations w/ Unbalanced Data

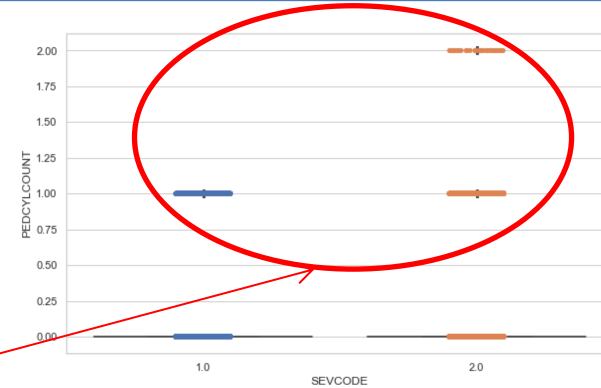
Unbalanced Data	X	Y	INTKEY	SEVCODE	SEVERITYI	PERSONC	PEDCOUN	PEDCYLC	VEHCOUN	SDOT_CO	INATTENT	UNDERIN	PEDROW	SPEEDING	HITPARKE	HofD	Year	Month	Day	colltype_	addrtype	junctype_	weather_	roadcond	lightcond
X	100%	-16%	1%	1%	1%	1%	1%	0%	-1%	1%	0%	0%	1%	0%	0%	1%	1%	0%	0%	1%	1%	-1%	-1%	-1%	-1%
Y	-16%	100%	3%	2%	2%	-1%	1%	3%	2%	-2%	-1%	-2%	2%	-3%	0%	2%	-2%	1%	0%	-3%	3%	-3%	2%	2%	3%
INTKEY	1%	3%	100%	20%	12%	7%	14%	8%	-5%	-4%	1%	-2%	15%	-5%	-4%	6%	9%	2%	2%	-46%	91%	-82%	0%	1%	1%
SEVCODE	1%	2%	20%	100%	80%	13%	24%	21%	-4%	18%	5%	7%	18%	1%	-4%	6%	8%	3%	2%	-14%	17%	-22%	-13%	-11%	-11%
SEVERITYDESC	1%	2%	12%	80%	100%	8%	19%	17%	-18%	15%	3%	3%	14%	-1%	-4%	-11%	-43%	-14%	-14%	6%	24%	7%	-7%	-6%	-5%
PERSONCOUNT	1%	-1%	7%	13%	8%	100%	-2%	-4%	38%	-15%	-7%	-16%	-2%	-9%	-3%	6%	2%	2%	2%	-2%	5%	-10%	20%	17%	21%
PEDCOUNT	1%	1%	14%	24%	19%	-2%	100%	-2%	-25%	27%	17%	27%	59%	-1%	-1%	4%	4%	2%	1%	7%	13%	-13%	-44%	-39%	-41%
PEDCYLCOUNT	0%	3%	8%	21%	17%	-4%	-2%	100%	-24%	40%	5%	6%	8%	-1%	-1%	4%	4%	2%	1%	-21%	7%	-9%	-12%	-13%	-8%
VEHCOUNT	-1%	2%	-5%	-4%	-18%	38%	-25%	-24%	100%	-40%	-15%	-30%	-19%	-16%	1%	18%	15%	8%	7%	-11%	-10%	-5%	42%	36%	42%
SDOT_COLCODE	1%	-2%	-4%	18%	15%	-15%	27%	40%	-40%	100%	24%	40%	26%	27%	-13%	-3%	0%	1%	0%	0%	-3%	-4%	-64%	-56%	-62%
INATTENTIONIND	0%	-1%	1%	5%	3%	-7%	17%	5%	-15%	24%	100%	24%	7%	3%	6%	1%	4%	1%	1%	-2%	-1%	0%	-33%	-30%	-30%
UNDERINFL	0%	-2%	-2%	7%	3%	-16%	27%	6%	-30%	40%	24%	100%	16%	26%	14%	3%	12%	2%	2%	-3%	-3%	7%	-65%	-57%	-69%
PEDROWNOTGRNT	1%	2%	15%	18%	14%	-2%	59%	8%	-19%	26%	7%	16%	100%	-1%	-1%	2%	1%	1%	1%	3%	13%	-13%	-32%	-28%	-29%
SPEEDING	0%	-3%	-5%	1%	-1%	-9%	-1%	-1%	-16%	27%	3%	26%	-1%	100%	0%	-3%	1%	1%	0%	-4%	-6%	4%	-30%	-21%	-37%
HITPARKEDCAR	0%	0%	-4%	-4%	-4%	-3%	-1%	-1%	1%	-13%	6%	14%	-1%	0%	100%	1%	5%	-1%	0%	1%	-4%	10%	-10%	-11%	-10%
HofD	1%	2%	6%	6%	-11%	6%	4%	4%	18%	-3%	1%	3%	2%	-3%	1%	100%	32%	8%	36%	-15%	0%	-14%	0%	-1%	1%
Year	1%	-2%	9%	8%	-43%	2%	4%	4%	15%	0%	4%	12%	1%	1%	5%	32%	100%	21%	22%	-24%	-10%	-36%	-7%	-6%	-7%
Month	0%	1%	2%	3%	-14%	2%	2%	2%	8%	1%	1%	2%	1%	1%	-1%	8%	21%	100%	7%	-8%	-4%	-11%	-2%	-2%	-3%
Day	0%	0%	2%	2%	-14%	2%	1%	1%	7%	0%	1%	2%	1%	0%	0%	36%	22%	7%	100%	-9%	-4%	-12%	-2%	-2%	-2%
colltype_code	1%	-3%	-46%	-14%	6%	-2%	7%	-21%	-11%	0%	-2%	-3%	3%	-4%	1%	-15%	-24%	-8%	-9%	100%	-36%	50%	5%	4%	4%
addrtype_code	1%	3%	91%	17%	24%	5%	13%	7%	-10%	-3%	-1%	-3%	13%	-6%	-4%	0%	-10%	-4%	-4%	-36%	100%	-68%	3%	3%	4%
junctype_code	-1%	-3%	-82%	-22%	7%	-10%	-13%	-9%	-5%	-4%	0%	7%	-13%	4%	10%	-14%	-36%	-11%	-12%	50%	-68%	100%	-4%	-5%	-5%
weather_code	-1%	2%	0%	-13%	-7%	20%	-44%	-12%	42%	-64%	-33%	-65%	-32%	-30%	-10%	0%	-7%	-2%	-2%	5%	3%	-4%	100%	92%	88%
roadcond_code	-1%	2%	1%	-11%	-6%	17%	-39%	-13%	36%	-56%	-30%	-57%	-28%	-21%	-11%	-1%	-6%	-2%	-2%	4%	3%	-5%	92%	100%	76%
lightcond_code	-1%	3%	1%	-11%	-5%	21%	-41%	-8%	42%	-62%	-30%	-69%	-29%	-37%	-10%	1%	-7%	-3%	-2%	4%	4%	-5%	88%	76%	100%

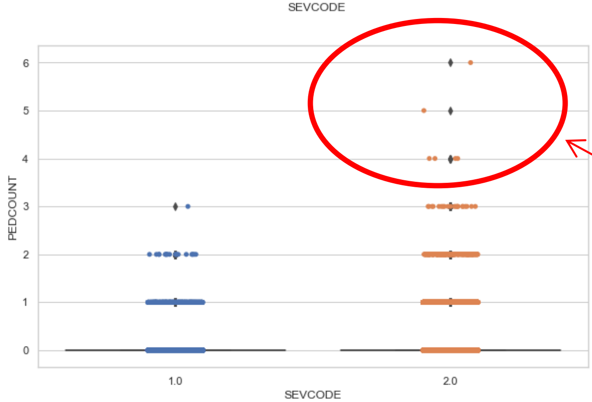
Data Pre-processing & Feature Engineering – Feature Correlations w/ Balanced Data

Balanced Data	X	Y	INTKEY	SEVCODE	SEVERITYI	PERSONC	PEDCOUN	PEDCYLC	VEHCOUN	SDOT_CO	INATTENT	UNDERIN	PEDROWI	SPEEDING	HITPARKE	HofD	Year	Month	Day	colltype_	addrtype	junctype_	weather_	roadcond	lightcond
X	100%	-16%	0%	1%	1%	1%	1%	0%	-1%	1%	0%	0%	1%	-1%	0%	2%	1%	0%	0%	1%	0%	-1%	-1%	-1%	-1%
Y	-16%	100%	4%	2%	2%	-1%	1%	3%	1%	-1%	0%	-1%	2%	-3%	0%	2%	-2%	1%	0%	-4%	4%	-4%	2%	1%	2%
INTKEY	0%	4%	100%	21%	16%	5%	16%	9%	-8%	-4%	1%	-1%	17%	-6%	-4%	6%	9%	1%	2%	-47%	93%	-84%	-1%	-1%	0%
SEVCODE	1%	2%	21%	100%	86%	14%	21%	19%	-4%	18%	5%	8%	15%	1%	-5%	6%	10%	3%	2%	-16%	19%	-25%	-13%	-12%	-11%
SEVERITYDESC	1%	2%	16%	86%	100%	10%	18%	16%	-14%	15%	4%	5%	13%	0%	-5%	-6%	-31%	-9%	-10%	-1%	23%	-2%	-9%	-8%	-7%
PERSONCOUNT	1%	-1%	5%	14%	10%	100%	-4%	-6%	41%	-16%	-7%	-16%	-4%	-9%	-3%	6%	1%	1%	1%	-2%	4%	-9%	20%	17%	21%
PEDCOUNT	1%	1%	16%	21%	18%	-4%	100%	-3%	-30%	30%	20%	32%	59%	-2%	-1%	4%	5%	2%	2%	10%	14%	-14%	-51%	-45%	-48%
PEDCYLCOUNT	0%	3%	9%	19%	16%	-6%	-3%	100%	-29%	45%	5%	7%	7%	-1%	-1%	4%	5%	2%	0%	-25%	8%	-10%	-14%	-14%	-9%
VEHCOUNT	-1%	1%	-8%	-4%	-14%	41%	-30%	-29%	100%	-44%	-17%	-32%	-22%	-15%	0%	14%	11%	6%	5%	-7%	-11%	-1%	45%	39%	45%
SDOT_COLCODE	1%	-1%	-4%	18%	15%	-16%	30%	45%	-44%	100%	25%	44%	29%	25%	-10%	-1%	1%	1%	0%	0%	-3%	-2%	-69%	-61%	-65%
INATTENTIONIND	0%	0%	1%	5%	4%	-7%	20%	5%	-17%	25%	100%	24%	8%	3%	4%	2%	5%	1%	1%	-1%	0%	0%	-33%	-30%	-31%
UNDERINFL	0%	-1%	-1%	8%	5%	-16%	32%	7%	-32%	44%	24%	100%	19%	25%	12%	4%	13%	2%	2%	-1%	-2%	6%	-65%	-57%	-69%
PEDROWNOTGRNT	1%	2%	17%	15%	13%	-4%	59%	7%	-22%	29%	8%	19%	100%	-1%	-1%	2%	1%	1%	1%	4%	16%	-15%	-37%	-32%	-35%
SPEEDING	-1%	-3%	-6%	1%	0%	-9%	-2%	-1%	-15%	25%	3%	25%	-1%	100%	0%	-3%	0%	1%	0%	-3%	-7%	6%	-28%	-19%	-35%
HITPARKEDCAR	0%	0%	-4%	-5%	-5%	-3%	-1%	-1%	0%	-10%	4%	12%	-1%	0%	100%	1%	4%	0%	0%	1%	-4%	9%	-8%	-8%	-9%
HofD	2%	2%	6%	6%	-6%	6%	4%	4%	14%	-1%	2%	4%	2%	-3%	1%	100%	31%	7%	35%	-14%	1%	-13%	-1%	-1%	0%
Year	1%	-2%	9%	10%	-31%	1%	5%	5%	11%	1%	5%	13%	1%	0%	4%	31%	100%	18%	19%	-21%	-8%	-33%	-6%	-5%	-6%
Month	0%	1%	1%	3%	-9%	1%	2%	2%	6%	1%	1%	2%	1%	1%	0%	7%	18%	100%	6%	-7%	-3%	-9%	-2%	-2%	-3%
Day	0%	0%	2%	2%	-10%	1%	2%	0%	5%	0%	1%	2%	1%	0%	0%	35%	19%	6%	100%	-7%	-3%	-11%	-2%	-2%	-2%
colltype_code	1%	-4%	-47%	-16%	-1%	-2%	10%	-25%	-7%	0%	-1%	-1%	4%	-3%	1%	-14%	-21%	-7%	-7%	100%	-39%	51%	2%	2%	1%
addrtype_code	0%	4%	93%	19%	23%	4%	14%	8%	-11%	-3%	0%	-2%	16%	-7%	-4%	1%	-8%	-3%	-3%	-39%	100%	-73%	1%	2%	2%
junctype_code	-1%	-4%	-84%	-25%	-2%	-9%	-14%	-10%	-1%	-2%	0%	6%	-15%	6%	9%	-13%	-33%	-9%	-11%	51%	-73%	100%	-2%	-3%	-3%
weather_code	-1%	2%	-1%	-13%	-9%	20%	-51%	-14%	45%	-69%	-33%	-65%	-37%	-28%	-8%	-1%	-6%	-2%	-2%	2%	1%	-2%	100%	92%	89%
roadcond_code	-1%	1%	-1%	-12%	-8%	17%	-45%	-14%	39%	-61%	-30%	-57%	-32%	-19%	-8%	-1%	-5%	-2%	-2%	2%	2%	-3%	92%	100%	76%
lightcond_code	-1%	2%	0%	-11%	-7%	21%	-48%	-9%	45%	-65%	-31%	-69%	-35%	-35%	-9%	0%	-6%	-3%	-2%	1%	2%	-3%	89%	76%	100%

Exploratory Data Analysis - Box Plots of Features by SEVERITY Class

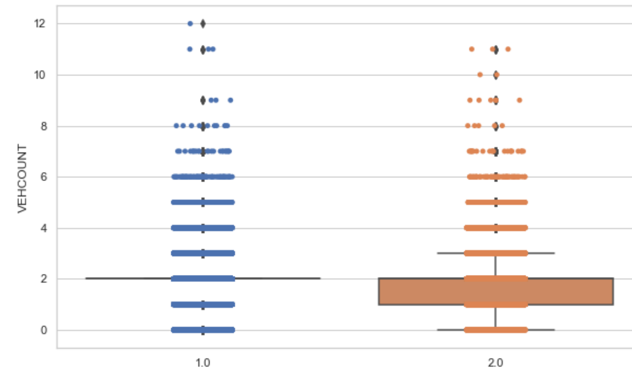


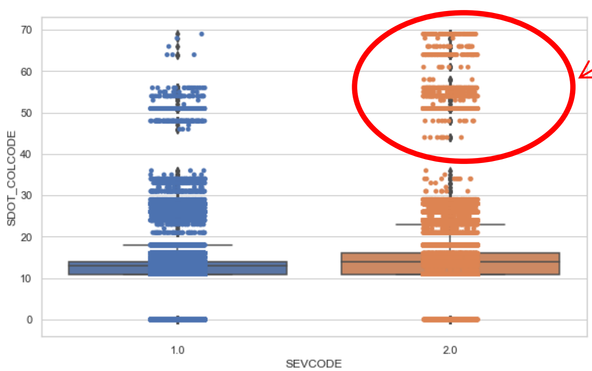


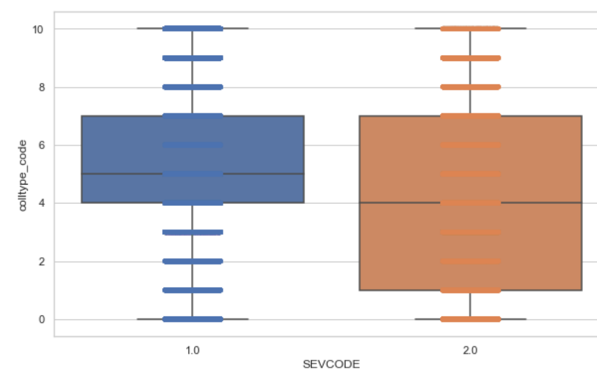


SEVERITY-2 Accidents are more likely to involve:

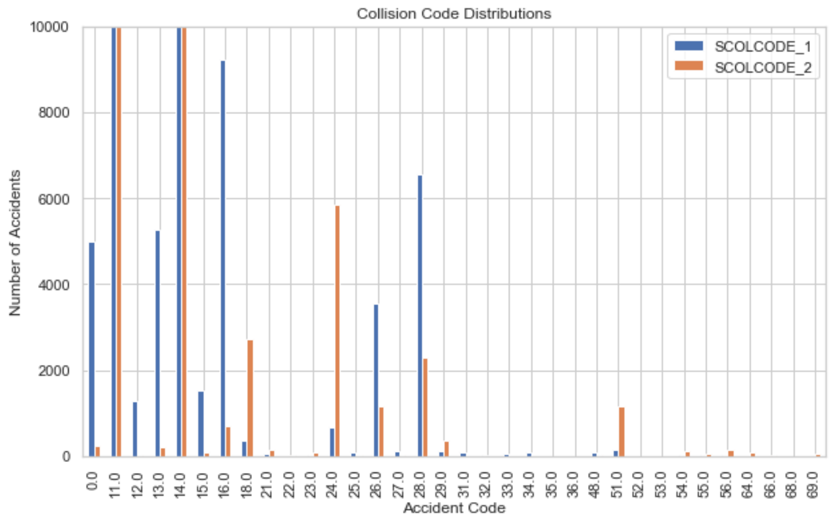
- More Pedestrians
- Cyclists
- More Lethal exogenous factors



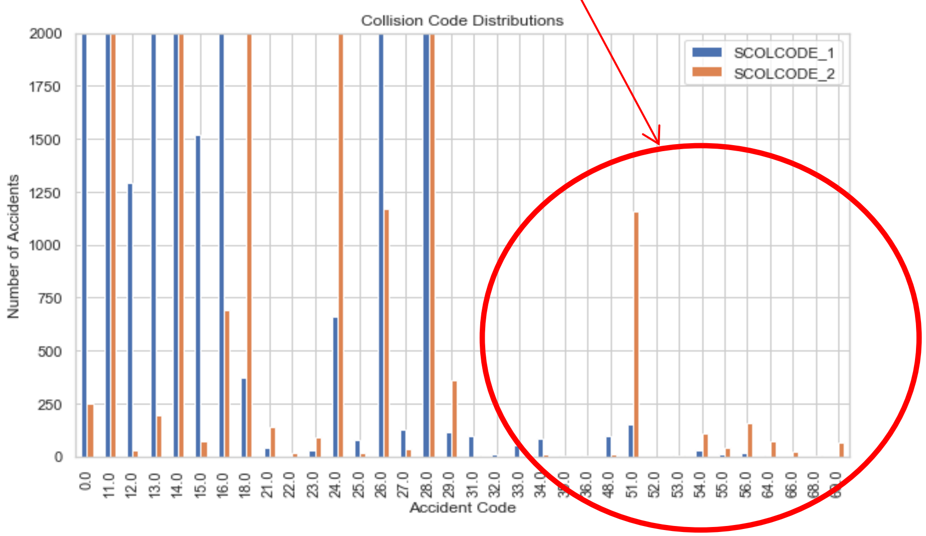




Exploratory Data Analysis – Histograms of SEVERITY DESCRIPTION by SEVERITY Class

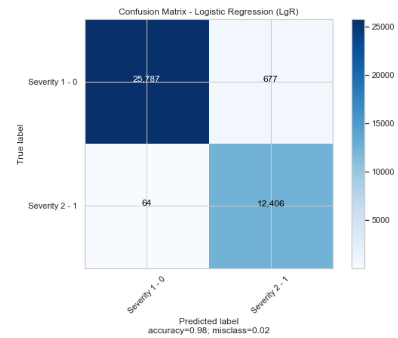


Level-2 accident distributions exhibit more codes with higher numbers that are related to accidents involving factors should reasonably correlate to more serious injury and fatalities including: collisions with heavy machinery, overturned vehicles, collisions with animals, vehicle fires, striking fixed objects, and collisions with trains.

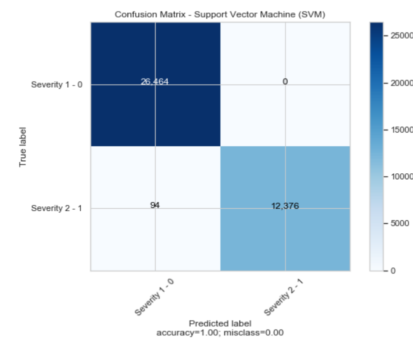


Classification model Objects and Confusion Matrices

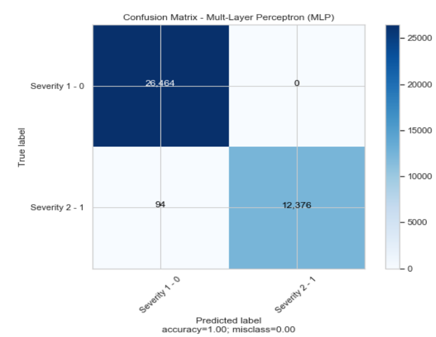
```
# Train a logistic regression classifier with default parameters using X_train and y_train.
# For the logistic regression classifier, create a precision recall curve and a roc curve
#using y_test and the probability estimates for X_test (probability it is fraud).
# Looking at the precision recall curve, what is the recall when the precision is `0.75`?
# Looking at the roc curve, what is the true positive rate when the false positive rate is `0.16`?
# *This function should return a tuple with two floats, i.e. `(recall, true positive rate)`.*

logreg = LogisticRegression(C=0.01, solver='liblinear',max_iter = 500)

logreg.fit(X_train, y_train)

y_pred = logreg.predict(X_val)
print('Accuracy of logistic regression classifier on test set: {:.2f}'.format(logreg.score(X_val, y_val)))
```




```
# Using X_train, X_test, y_train, y_test (as defined above), train a SVC classifer using the default parameters.
# What is the accuracy, recall, and precision of this classifier?
from sklearn.metrics import precision_recall_curve
from sklearn.metrics import plot_precision_recall_curve
import matplotlib.pyplot as plt
from sklearn.metrics import average_precision_score

# *This function should a return a tuple with three floats, i.e. `(accuracy score, recall score, precision score)`.*
#def SVC_classifier():

svm = SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
              decision_function_shape='ovr', degree=3, gamma='scale',
              kernel='rbf', max_iter=-1, probability=False, random_state=None,
              shrinking=True, tol=0.001, verbose=1).fit(X_train, y_train)
y_pred = svm.predict(X_val)
accuracy_sc = svm.score(X_val, y_val)
recall_sc = recall_score(y_val, y_pred)
precision_sc = precision_score(y_val, y_pred)
average_precision = average_precision_score(y_val, y_pred)
disp = plot_precision_recall_curve(svm, X_val, y_val)
disp.ax_.set_title('2-class Precision-Recall curve: '
                   'AP={0:0.2f}'.format(average_precision))
print('Average precision-recall score: {0:0.2f}'.format(
      average_precision))
```




```
model3 = MLPClassifier(hidden_layer_sizes=(27,27,10),
              activation='relu',
              solver='adam',
              learning_rate='adaptive',
              early_stopping=True,
              max_iter=500, alpha=0.0001,
              verbose=0,  random_state=21,)

model3.fit(X_train, y_train)
y_pred = model3.predict(X_val)
test_acc = accuracy_score(y_val, y_pred) * 100.
loss_values = model3.loss_curve_
```



Classification model performance comparison summary

SEVERITYDESC feature included							
Unbalanced Training/Validation			precision	recall	f1-score	support	Accuracy
MLP	Level 1	0	1.00	1.00	1.00	22030	1.00
	Level 2	1	1.00	0.99	1.00	9118	
LogReg	Level 1	0	1.00	0.98	0.99	22030	0.99
	Level 2	1	0.96	0.99	0.98	9118	
SVM	Level 1	0	1.00	1.00	1.00	22030	1.00
	Level 2	1	1.00	0.99	1.00	9118	

Confusion Matrix

MLP	True Labels	Severity 1	71%	0%
		Severity 2	0%	29%
LogReg		Severity 1	70%	1%
		Severity 2	0%	29%
SVM		Severity 1	71%	0%
		Severity 2	0%	29%
			Severity 1	Severity 2
			Predicted Labels	

SEVERITYDESC feature included							
Balanced Training/Validation			precision	recall	f1-score	support	Accuracy
MLP	Level 1	0	0.99	1.00	1.00	9430	1.00
	Level 2	1	1.00	0.99	1.00	9191	
LogReg	Level 1	0	1.00	0.97	0.99	9430	0.99
	Level 2	1	0.97	1.00	0.99	9191	
SVM	Level 1	0	0.99	1.00	1.00	9430	1.00
	Level 2	1	1.00	0.99	1.00	9191	

Confusion Matrix

MLP	True Labels	Severity 1	51%	0%
		Severity 2	0%	49%
LogReg		Severity 1	49%	1%
		Severity 2	0%	49%
SVM		Severity 1	51%	0%
		Severity 2	0%	49%
			Severity 1	Severity 2
			Predicted Labels	

- MLP model overall outperformed Logistic Regression and Support Vector Machine
- Balancing the data set did not result in materially improved performance

Unbalanced Test Data			precision	recall	f1-score	support	Accuracy
MLP	Level 1	0	0.77	0.95	0.85	22030	0.76
	Level 2	1	0.70	0.30	0.42	9118	
LogReg	Level 1	0	0.74	0.98	0.84	22030	0.74
	Level 2	1	0.80	0.15	0.26	9118	
SVM	Level 1	0	0.74	0.99	0.85	22030	0.75
	Level 2	1	0.90	0.18	0.30	9118	

Confusion Matrix

MLP	True Labels	Severity 1	67%	4%
		Severity 2	20%	9%
LogReg		Severity 1	67%	4%
		Severity 2	20%	9%
SVM		Severity 1	70%	1%
		Severity 2	24%	5%
			Severity 1	Severity 2
			Predicted Labels	

Balanced Test Data			precision	recall	f1-score	support	Accuracy
MLP	Level 1	0	0.75	0.60	0.66	9232	0.70
	Level 2	1	0.67	0.81	0.73	9389	
LogReg	Level 1	0	0.62	0.68	0.65	9232	0.63
	Level 2	1	0.65	0.59	0.62	9389	
SVM	Level 1	0	0.63	0.76	0.69	9232	0.66
	Level 2	1	0.70	0.57	0.63	9389	

Confusion Matrix

MLP	True Labels	Severity 1	30%	20%
		Severity 2	10%	41%
LogReg		Severity 1	34%	16%
		Severity 2	21%	30%
SVM		Severity 1	37%	12%
		Severity 2	22%	29%
			Severity 1	Severity 2
			Predicted Labels	

- Omission of the encoded SEVERITYDESC feature resulted in a 25-30% performance drop for all three models using both unbalanced and balanced data

Conclusion and model refinements

Project Deliverable to design and implement useful predictive machine learning models to classify traffic accident severity class achieved:

- Highly accurate models generalized well in both validation and test sets
- MLP model chosen as best in class based on:
    - ✓ low false positive rate
    - ✓ high accuracy
    - ✓ performance speed

Possible refinements include the development of additional features that are more correlated to SEVERITYDESC:

- Ideas for additional/better features include:
    - Additional features on more specific number and type of injuries
    - Better fatality feature data which was rendered unusable due to large number of null values